Fundamentals of Modeling, Data Assimilation, and High-performance Computing

Lecture Notes

Richard B. Rood

Introduction

This lecture will introduce the concepts of modeling, data assimilation and high-performance computing as it relates to the study of atmospheric composition. The lecture will work from basic definitions and will strive to provide a framework for thinking about development and application of models and data assimilation systems. It will not provide technical or algorithmic information, leaving that to textbooks, technical reports, and ultimately scientific journals. References to a number of textbooks and papers will be provided as a gateway to the literature.

The text will be divided into four major sections.

- Modeling
- Data Assimilation
- Observing System
- High-performance Computing

Modeling

Dictionary definitions of model include:

- A work or construction used in testing or perfecting a final product.
- A schematic description of a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further studies of its characteristics.

In atmospheric modeling the scientist is generally faced with a set of observations of parameters, for instance, wind, temperature, water, ozone, *etc.*, as well as either the knowledge or expectation of correlated behavior between the different parameters. A number of types of models could be developed to describe the observations. These include:

- Conceptual or heuristic models which outline in the simplest terms the processes that describe the interrelation between different observed phenomena. These models are often intuitively or theoretically based. An example would be the tropical pipe model of Plumb and Ko [1992], which describes the transport of long-lived tracers in the stratosphere.

- Statistical models which describe the behavior of the observations based on the observations themselves. That is the observations are described in terms of the mean, the variance, and the correlations of an existing set of observations. Johnson *et al.* [2000] discuss the use of statistical models in the prediction of tropical sea surface temperatures.

- Physical models which describe the behavior of the observations based on first principle tenets of physics (chemistry, biology, *etc.*). In general, these principles are expressed as mathematical equations, and these equations are solved using discrete numerical methods. Good introductions to modeling include [Trenberth, 1992; Jacobson, 1998; Randall, 2000]

In the study of geophysical phenomena, there are numerous sub-types of models. These include comprehensive models which attempt to model all of the relevant couplings or interactions in a system and mechanistic models which have one or more parameters prescribed, for instance by observations, and then the system evolves relative to the prescribed parameters. All of these models have their place in scientific investigation, and it is often the interplay between the different types and sub-types of models that leads to scientific advance.

Models are used in two major roles. The first role is diagnostic, in which the model is used to determine and to test the processes that are thought to describe the observations. In this case, it is determined whether or not the processes are well known and adequately described. In general, since models are an investigative tool, such studies are aimed at determining the nature of unknown or inadequately described processes. The second role is prognostic; that is, the model is used to make a prediction. All types of models can be used in these roles.

In all cases the model represents a management of complexity; that is, the scientist is faced with a complex set of observations and their interactions and is trying to manage those observations in order to develop a quantitative representation. In the case of physical models, which are implicitly at the focus of this lecture, a comprehensive model would represent the cumulative knowledge of the physics (chemistry, biology, *etc.*) that describe the observations. It is tacit, that an accurate, comprehensive physical ~~model is the most robust way to forecast; that is, to predict the future.~~

While models are a scientist's approach to manage and to probe complex systems, today's comprehensive models are themselves complex. In fact, the complexity and scope of models is great enough that teams of scientists are required to contribute to modeling activities. Two consequences of complexity of models are realized in computation and interpretation. Comprehensive models of the Earth system remain outside the realm of the most capable computational systems. Therefore, the problem is reduced to either looking at component models of the earth system, *i.e.,* atmosphere, ocean, land, cryosphere, lithosphere, or at models where significant compromises are taken in the representation of processes in order to make them computationally feasible.

More challenging, and in fact the most challenging aspect of modeling, is the interpretation of model results. It is much easier to build models than it is to do quantitative evaluation of models and observations.

In order to provide an overarching background for thinking about models it is useful to consider the elements of the modeling, or simulation, framework described in Figure 1. In this framework are six major ingredients. The first are the boundary and initial conditions. For an atmospheric model, boundary conditions are topography and sea surface temperature; boundary conditions are generally prescribed from external sources of information. It is the level of prescription of boundary conditions and the conventions of the particular discipline that determine whether or not a model is termed mechanistic.

The next three items in the figure are intimately related. They are the representative equations, the discrete or parameterized equations, and the constraints drawn from theory. The representative equations are the analytic form of forces or factors that are considered important in the representation of the evolution of a set of parameters. In general, all of the analytic expressions used in atmospheric modeling are approximations; therefore, even the equations the modeler is trying to solve have *a priori* errors. Generally in the construction of a model, only terms that are expected to be important are included in the analytic expressions; that is, the equations are scaled from some more complete representation [see Holton, 2004]. The solution is, therefore, a balance amongst competing forces and tendencies. Most commonly, the analytic equations are a set of non-linear partial differential equations.

The discrete or parameterized equations arise because it is generally not possible to solve the analytic equations in closed form. The strategy used by scientists is to develop a discrete representation of the equations which are then solved using numerical techniques. These solutions are, at best, discrete estimates to solutions of the analytic equations. The discretization and parameterization of the analytic equations introduce a large source of error. This introduces another level of balancing in the model; namely, these errors are generally managed through a subjective balancing process that keeps the numerical solution from running away to obviously incorrect estimates.

While all of the terms in the analytic equation are potentially important, there are conditions or times when there is a dominant balance between, for instance, two terms. An example of this is thermal wind balance in the middle latitudes of the stratosphere [see Holton, 2004]. It is these balances, generally at the extremes of spatial and temporal scales, which provide the constraints drawn from theory. Such constraints are generally involved in the development of conceptual or heuristic models. If the modeler implements discrete methods which consistently represent the relationship between the analytic equations and the constraints drawn from theory, then the modeler maintains a substantive scientific basis for the interpretation of model results.

The last two items in Figure 1 represent the products that are drawn from the model. These are divided into two types: primary products and derived products. The primary products are variables such as wind, temperature, water, ozone – parameters that are most often, explicitly modeled; that is, an equation is written for them. The derived products are often functional relationships between the primary products; for instance, potential vorticity. A common derived product is the balance, or the budget, of the different terms of the discretized equations. The budget is then studied, explicitly, on how the balance is maintained and how this compares with budgets derived directly from observations. In general, the primary products can be directly evaluated with observations and errors of bias and variability estimated. If attention has been paid in the discretization of the analytic equations to honor the theoretical constraints, then the derived products will behave consistently with the primary products and theory. They will have errors of bias and variability, but they will behave in a way that supports scientific scrutiny.

In order to explore the elements of the modeling framework described above, the following continuity equation for a constituent, A, will be posed as the representative equation. The continuity equation represents the conservation of mass for a constituent and is an archetypical equation of geophysical models. Brasseur and Solomon [1986] and Dessler [2000] provide good backgrounds for understanding atmospheric chemistry and transport. The continuity equation for A is:

$$\partial A/\partial t = -\nabla \bullet \mathbf{U}A + M + P - LA - n/HA + q/H \qquad (1)$$

Where,

- A is some constituent
- **U** is velocity → "resolved" transport, "advection"
- M is "Mixing" → "unresolved" transport, parameterization
- P is production
- L is loss
- n is "deposition velocity"
- q is emission
- H is representative length scale for n and q
- t is time
- ∇ is the gradient operator

Attention will be focused on the discretization of the resolved advective transport. Figures 2 and 3 illustrate the basic concepts. On the left of the figure a mesh has been laid down to cover the spatial domain of interest. In this case it is a rectangular mesh. The mesh does not have to be rectangular, uniform, or orthogonal. In fact the mesh can be unstructured or can be built to adapt to the features that are being modeled. The choice of the mesh is determined by the modeler and depends upon the diagnostic and prognostic applications of the model [see Randall, 2000]. The choice of mesh can also be determined by the computational advantages that might be realized.

Points can be prescribed to determine location with the mesh. In Figure 2 both the advective velocity and the constituent are prescribed at the center of the cell. In

Figure 3, the velocities are prescribed at the center of the cell edges, and the constituent is prescribed in the center of the cell. There are no hard and fast rules about where the parameters are prescribed, but small differences in their prescription can have huge impact on the quality of the estimated solution to the equation, *i.e.* the simulation. The prescription directly impacts the ability of the model to represent conservation properties and to provide the link between the analytic equations and the theoretical constraints [see Rood, 1987; Lin 2004]. In addition, the prescription is strongly related to the stability of the numerical method; that is, the ability to represent any credible estimate at all.

A traditional and intuitive approach to discretization is to use differences calculated across the expanse of the grid cell to estimate partial derivatives. This is the foundation of the finite-difference method, and finite-differences appear in one form or another in various components of most models. Differences can be calculated from a stencil that covers a single cell or weighted values from neighboring cells can be used. From a numerical point of view, the larger the stencil, the more cells that are used, the more accurate the approximation of the derivative. Spectral methods, which use orthogonal expansion functions to estimate the derivatives, essentially use information from the entire domain. While the use of a large stencil generally increases the accuracy of the estimate of the partial derivatives, it also increases the computational cost and means that discretization errors are correlated across large portions of the domain.

The use of numerical techniques to represent the partial differential equations that represent the model physics is a straightforward way to develop a model. However, there are many approaches to discretization of the dynamical equations that govern geophysical processes [Jacobson, 1998; Randall, 2000]. Given that these equations are, in essence, shared by many scientific disciplines, there are sophisticated and sometimes similar developments in many different fields. One approach that has been recently adopted by several modeling centers is described in Lin [2004]. In this approach the cells are treated as finite volumes and piecewise continuous functions are fit locally to the cells. These piecewise continuous functions are then integrated around the volume to yield the forces acting on the volume. This method, which was derived with physical consistency as a requirement for the scheme, has proven to have numerous scientific advantages. The scheme uses the philosophy that if the physics are properly represented, then the accuracy of the scheme can be robustly built on a physical foundation. In addition, the scheme, which is built around local stencils, has numerous computational advantages.

Douglass *et al.* [2003] and Schoeberl *et al.* [2003] have demonstrated the improved quality that follows from implementation of the finite volume scheme. In their studies they investigate the transport and mixing of atmospheric constituents in the upper troposphere and the lower stratosphere. Through a combination of analysis of observations, a hierarchy of models, and the relationship to theoretical constraints, they demonstrate that both the horizontal and vertical transport is better represented with the finite volume scheme. Further, their comparisons of experiments using winds from several data assimilation systems to calculate transport establish the importance of physical consistency in the representation of budgets of the constituent continuity equation.

Data Assimilation

The definition of assimilation from the dictionary is:

- To incorporate or absorb; for instance, into the mind or the prevailing culture

For Earth science, assimilation is the incorporation of observational information into a physical model. Or more specifically:

- Data assimilation is the objective melding of observed information with model-predicted information.

Returning to the discussion of model types in the previous section, assimilation rigorously combines statistical modeling with physical modeling; thus, formally connecting the two approaches. Daley [1991] is the standard text on data assimilation. Cohn [1997] explores the theory of data assimilation and its foundation in estimation theory. Swinbank *et al.* [2003] is a collection of tutorial lectures on data assimilation. Assimilation is difficult to do well, easy to do poorly, and its role in Earth science is expanding and sometimes controversial.

Figure 4 shows elements of an assimilation framework that parallels the modeling elements in Figure 1. The concept of boundary conditions remains the same; that is, some specified information at the spatial and temporal domain edges. Of particular note, the motivation for doing data assimilation is often to provide the initial conditions for predictive forecasts.

Data assimilation adds an additional forcing to the representative equations of the physical model; namely, information from the observations. This forcing is formally added through a correction to the model that is calculated, for example, by [see Stajner *et al.,* 2001]:

$$(\boldsymbol{O}P_f\boldsymbol{O}^T + R)x = A_o - \boldsymbol{O}A_f \qquad (2)$$

The terms in the equation are as follows:

- A_o are observations of the constituent
- A_f are model forecast, simulated, estimates of the constituent
- $\boldsymbol{O}$ is the observation operator
- P_f is the error covariance function of the forecast
- R is the error covariance function of the observations
- x is the innovation that represents the observation-based correction to the model
- ^{T} is the matrix transform operation

The observation operator, ***O***, is a function that maps the parameter to be assimilated onto the spatial and temporal structure of the observations. In its simplest form, the observation operator is an interpolation routine. Often, however, it is best to perform assimilation in observation space, and in the case of satellite observations the measurements are radiances. Therefore, the observation operator might include a forward radiative transfer calculation from the model's geophysical parameters to radiance space. While this is formally robust, in practice, it is sometimes less than successful because of loss of intuitiveness and computational problems. Therefore, initial experiments with assimilation are often most productively undertaken using retrieved geophysical parameters.

The error covariance functions, P_f and R, represent the errors, respectively, of the information from the forecast model and the information from the observations. This explicitly shows that data assimilation is the error-weighted combination of information from two primary sources. These error covariance functions are generally not well known. From first principles, the error covariance functions are prohibitive to calculate. Hence, models are generally developed to represent the error covariance functions. Stajner *et al.* [2001] show a method for estimating the error covariances in an ozone assimilation system.

Parallel to the elements in the simulation framework (Figure 1), discrete numerical methods are needed to estimate the errors as well as to solve the matrix equations in Equation (2). How and if physical constraints from theory are addressed is a matter of both importance and difficulty. Often, for example, it is assumed that the increments of different parameters that are used to correct the model are in some sort of physical balance. For instance, wind and temperature increments might be expected to be in geostrophic balance. However, in general, the data insertion process acts like an additional physics term in the equation and contributes a significant portion of the budget. This, explicitly, alters the physical balance defined by the representative equations of the model. Therefore, there is no reason to expect that the correct geophysical balances are represented in an assimilated data product. This is contrary to the prevailing notion that the model and observations are 'consistent' with one another after assimilation.

The final two elements in Figure 4 are, again, the products. In a good assimilation the primary products, in general the prognostic variables, are well estimated. That is, ~~both the bias errors and the variance errors are reduced. However, the derived products~~ are likely to be physically inconsistent because of the nature of the corrective forcing added by the observations. These errors are often found to be larger than those in self-determining model simulations. Molod *et al.* [1996] and Kistler *et al.* [2001] provide discussions on the characteristics of the errors associated with primary and derived products in data assimilation systems. The nature of the errors described in these papers is consistent with errors in present-day assimilation systems.

A schematic of an assimilation system is given in Figure 5. This is a sequential assimilation system where a forecast is provided to a statistical analysis algorithm that calculates the merger of model and observational information. In this example, errors are specified based on external considerations and methods. There is a formal interface between the statistical analysis algorithm and the model prediction which performs a quality assessment of the information prior to the merger; this quality control algorithm will be discussed more fully below. The figure shows, explicitly, two input streams for the observations. The first of these streams represent the observations that will be assimilated with the model prediction. The other input stream represents observations that will not be assimilated. This second stream of observations could be, for example, a new type of observation whose error characteristics are being determined relative to the existing assimilation system. The second stream might also represent an ancillary data set that is being used in quality control decisions. This type of monitoring function finds many applications, and data assimilation systems are excellent tools for determining anomalies in input data streams.

In Figure 4 the products of the assimilation were classified as primary and derived estimates of geophysical parameters. The following classification of products describes the collective information from the data assimilation system. These are indicated in Figure 5 and listed below:

- Analysis: The analysis is the merged combination of model information and observational information. The analysis is the best estimate of the state of the system based on the optimization criteria and error estimates.

- Forecast/simulation: The forecast/simulation is a model run that starts from an initial condition defined by the analysis. For some amount of time this model run is expected to represent the state of the system with some deterministic accuracy. For this case the model run is a forecast. After a certain amount of time the model run is no longer expected to represent the particular state of the system; though, it might represent the average state. In this case the model run is simply a simulation that has been initialized with a realistic state estimate at some particular time.

- Observation minus forecast increment: The observation minus forecast increment, often the O-F, gives a raw estimate of the agreement of the forecast information with the observation information prior to assimilation. Usually, a small O-F increment indicates a high quality forecast, and O-F increments are used as a primary measure of the quality of the assimilation. O-F increments are exquisitely sensitive to changes in the system and are the primary quantity used for monitoring the stability and quality of the input data streams. Study of the O-F is useful for determining the spatial and temporal characteristics of some model errors.

- Observation minus analysis increment: The observation minus analysis increment represents the actual changes to the model forecast that are derived from the statistical analysis algorithm. Therefore, they represent in some bulk sense the error weighted impact of the O-F increments. If the assimilation system weighs the observations heavily relative to the forecast, then the observation minus analysis increments will have significant differences relative to the O-F increments. The opposite is also true; if the model information is weighed more heavily than the observational information then there will be little change to the O-F increments. If either of these extremes are realized the basic assumptions of the assimilation problem need to be reconsidered.

As suggested earlier, the specification of forecast and model error covariances and their evolution with time is a difficult problem. In order to get a handle on these problems it is generally assumed that the observational errors and model errors are unbiased over some suitable period of time, *e.g.* the length of the forecast between times of data insertion. It is also assumed that the errors are in a Gaussian distribution. The majority of assimilation theory is developed based on these assumptions, which are, in fact, not valid assumptions. In particular, when the observations are biased, there would the expectation that the actual balance of geophysical terms is different from the balance determined by the assimilation. Furthermore, since the biases will have spatial and temporal variability, the balances determined by the assimilation are quite complex. Aside from biases between the observations and the model prediction, there are biases between different observation systems of the same parameters. These biases are potentially correctible if there is a known standard of accuracy defined by a particular observing system. However, the problem of bias is a difficult one to address and perhaps the greatest challenge facing assimilation [see, Dee and da Silva, 1998]

As a final general consideration, there are many time scales represented by the representative equations of the model. Some of these time scales represent balances that are achieved almost instantly between different variables. Other time scales are long, important to, for instance, the general circulation which will determine the distribution of long-lived trace constituents. It is possible in assimilation to produce a very accurate representation of the observed state variables and those variables which are balanced on fast time scales. On the other hand, improved estimates in the state variables are found, at least sometimes, to be associated with degraded estimates of those features determined by long time scales. Conceptually, this can be thought of as the impact of bias propagating through the physical model. With the assumption that the observations are fundamentally accurate, this indicates errors in the specification of the physics that demand further research.

~~Data assimilation has had dramatic impacts in the improvement of weather~~ forecasts. In other applications the benefits of assimilation have been more difficult to realize. Therefore, scientists need to determine the appropriateness of assimilation or using assimilated data products in their studies. The list below provides goals of the assimilation of ozone data. These goals are examples which can be extended to the assimilation of other geophysical parameters.

- Mapping: There are spatial and temporal gaps in the ozone observing system. A basic goal of ozone assimilation is to provide vertically resolved global maps of ozone.

- Short-term ozone forecasting: There is interest in providing operational ozone forecasts in order to predict the fluctuations of ultraviolet radiation at the surface of the earth [Long et al., 1996].
- Chemical constraints: Ozone is important in many chemical cycles. Assimilation of ozone into a chemistry model provides constraints on other observed constituents and helps to provide estimates of unobserved constituents.
- Unified ozone data sets: There are several sources of ozone data with significant differences in spatial and temporal characteristics as well as their expected error characteristics. Data assimilation provides a potential strategy for combining these data sets into a unified data set.
- Tropospheric ozone: Most of the ozone is in the stratosphere, and tropospheric ozone is sensitive to surface emission of pollutants. Therefore, the challenges of obtaining accurate tropospheric ozone measurements from space are significant. The combination of observations with the meteorological information provided by the model offers one of the better approaches available to obtain global estimates of tropospheric ozone.
- Improvement of wind analysis: The photochemical time scale for ozone is long compared with transport timescales in the lower stratosphere and upper troposphere. Therefore ozone measurements contain information about the wind field what might be obtained in multi-variate assimilation.
- Radiative transfer: Ozone is important in both longwave and shortwave radiative transfer. Therefore accurate representation of ozone is important in the radiative transfer calculations needed to extract (retrieve) information from many satellite instruments. In addition, accurate representation of ozone has the potential to impact the quality of the temperature analysis in multi-variate assimilation.
- Observing system monitoring: Ozone assimilation offers an effective way to characterize instrument performance relative to other sources of ozone observations as well as the stability of measurements over the lifetime of an instrument.
- Retrieval of ozone: Ozone assimilation offers the possibility of providing more accurate initial guesses for ozone retrieval algorithms than are currently available.
- Assimilation research: Ozone (constituent) assimilation can be productively approached as a univariate linear problem. Therefore it is a good framework for investigating assimilation science; for example, the impact of flow dependent covariance functions.
- Model and observation validation: Ozone assimilation provides several approaches to contribute to the validation of models and observations.

Some of the goals mentioned above can be meaningfully addressed with the current state of the art. Others cannot. It is straightforward to produce global maps of total column ozone which can be used in, for instance, radiative transfer calculations. The use of ozone measurements to provide constraints on other reactive species is an application that has been explored since the 1980's [see, Jackman *et al.*, 1987] and modern data assimilation techniques potentially advance this field. The impact of ozone assimilation on the meteorological analysis of temperature and wind, and hence improvement of the weather forecast, is also possible. The most straightforward impact would be on the temperature analysis in the stratosphere. The improvement of the wind analysis is a more difficult challenge and confounded by the fact that where improvements in the wind analyses are most needed, the tropics, the ozone gradients are relatively weak. The use of ozone assimilation to monitor instrument performance and to characterize new observing systems is currently possible and productive [see, Stajner *et al.,* 2004]. The improvement of retrievals using assimilation techniques to provide ozone first guess fields that are representative of the specific environmental conditions is also an active research topic. The goal of producing unified ozone data sets from several instruments is of little value until bias can be correctly accommodated in data assimilation. This final topic will be discussed more fully below.

There has been much written in the assimilation literature about the various approaches to the assimilation algorithm and the use of assimilated data sets in many types of applications. There has been relatively little written about the quality control of the observations; that is, the interface between the observations and the model predicted data sets. The successes or failures of data assimilation systems can however be directly related to decisions made in the quality control [see, Dee *et al.,* 2001]. A simple description of quality control is given in Figure 6.

On the left side of the figure are three sources of observational information, two satellites and one non-satellite source. These three sources of observations might represent a nadir viewing temperature sounder, a limb viewing temperature sounder, and radiosonde temperatures. Even if perfectly accurate, each of these observing systems would provide different measurements because of the sampling characteristics of the instrument. Quality control of the observations might proceed as follows. Each type of observation is likely to come with some information about the anticipated quality of the data. This information might indicate whether or not an observation is far outside the expected value based on the previous performance of the instrument, or alternatively, that the instrument was in an observing mode known to have low reliability. Further investigation of observations that are flagged, as say, suspicious, might reveal that there is a region of geographical consistency; that is, a region of suspicious data. This region could represent a meaningful geophysical feature, perhaps a developing storm, or it might represent a regional contamination of the observations, perhaps clouds or an erupting volcano.

~~If the investigation of the observation suggests that the observations might be of~~ geophysical interest, then intercomparison with other types of observations can confirm this suggestion. Since the different types of observations might have different environmental sensitivities, the identification of a regional anomaly in all of the observation types would add weight to favor the inclusion of the suspicious data in the assimilation. Finally, the model prediction can be brought into the decision making process. The model is an estimate of the projected value of the observation, and the observation minus forecast information is a sensitive indicator of information. If the model suggests there is a developing storm, then inclusion of the data is likely to better represent the forecast of that storm. If the model does not show a storm, but all three types of observations suggest that there is a storm, then the analysis will reflect the storm, and an otherwise missed feature will be correctly forecast.

Quality control decisions are difficult and can have significant impact on the assimilation quality [Dee *at al.*, 2001]. It is intuitive that a handful of observations taken near clouds that represent a real developing storm will have much more impact than additional observations in clear skies where persistence is expected. In a scientific investigation of the data system, the data rejected by the quality control demand further investigation. They could reflect an instrument malfunction or an operator or transmission error. Another possibility is that the field of view is contaminated by a cloud, or perhaps a volcanic eruption has been detected. Finally, systematic rejection of data might suggest that the assimilation system is drifting because the error covariances are not robustly specified, or that a new geophysical phenomenon, perhaps a trend, is being measured.

Figure 7 shows an example from an assimilation of ozone data [Wargan *et al.* 2005]. In this example, there are two satellite instruments, the Solar Backscattered Ultraviolet/2 (SBUV) and the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS). SBUV is a nadir sounder and measures very thick layers with the vertical information in the middle and upper stratosphere. MIPAS is a limb sounder with much finer vertical resolution and measurements extending into the lower stratosphere. SBUV also measures total column ozone, which is assimilated in all experiments. The results from three assimilation experiments are shown through comparison with an ozonesonde profile. Ozonesondes were not assimilated into the systems; therefore, these data provide an independent measure of performance.

There are several attributes to be noted in Figure 7. The quality of the MIPAS-only (+ SBUV total column) assimilation is the best of those presented. This suggests that the vertical resolution of MIPAS instrument is having a large impact. Even though the MIPAS observations are assimilated only above 70 hPa, the assimilation captures the essence of the structure of the ozone profile down to 300 hPa. This indicates that the model information in the lower stratosphere and upper troposphere is geophysically meaningful. Further, the model is effectively distributing information in the horizontal between the satellite profiles. The comparison with the SBUV-only (+ SBUV total column) assimilation shows that the thick-layered information of the SBUV observations, ~~even in combination with the model information, does not represent the ozone peak very~~ well. This impacts the quality of the lower stratospheric analysis as the column is adjusted to represent the constraints of the total ozone observations. Finally, from first principles, the combined SBUV and MIPAS assimilation might be expected to be the best; this should have the maximum amount of information. This is not found to be the case, and suggests that the weights of the various error covariances and the use of the observations can be improved. The optimal balance of nadir and limb observations is not straightforward, and these experiments reveal the challenges that need to be addressed when multiple types of instruments are used in data assimilation.

The Observing System

The example of ozone assimilation discussed in Figure 7 hints at the attention that needs to be paid to the characteristics of the instruments in the observing system. There are questions of accuracy and bias between different observations of the same parameter. Similarly, there are questions of how different instruments measure the variability of geophysical parameters. When different, but correlated, parameters are measured, for example, ozone and temperature, there is the question of how measurement errors fit into the correlated behavior of the parameters.

Data assimilation brings yet another level of attention to the observations; namely, the specific characteristics of the observing system: How do the footprints of two instruments impact their use? What is depth of the averaging kernel? How are limb scanning and nadir measurements used together? How are sparse, high quality localized observations used? Should dense observations be sampled or averaged prior to use? How is information from integrated quantities, such as outgoing longwave radiation, used? One approach to answering these questions is to transform the model variables into the same space as the observations through the observation operator (See Equation 2). As noted, above, the rigor of this strategy is often challenged by the reality of the implementation and the interpretation of the problem so that the approach is not straightforward or successful.

The observational data used in data assimilation are often broken into three categories, listed below:

- Conventional Data: Conventional data are those data that are, in principle, from the pre-satellite era. More generally, conventional data are non-satellite data. These data include surface observations and balloon soundings, as well as ship and aircraft observations. Many of these data are collected operationally, and they are a critical part of meteorological (and oceanographic and land-surface) assimilation systems. There are also data collected in research missions that are of exceptional quality that find their use in data assimilation, often as independent validation data. Though non-satellite data might be classified as conventional, the optimal use of these observations requires careful attention to the details of the data systems. Lait [2002] provides an interesting examination of the radiosonde network and the impact that instruments from different countries and manufacturers have on the quality of the analysis.

- Operational Satellite Data: Operational satellite data are those data taken routinely to support national weather prediction centers. These data are taken and processed in real-time and distributed around the world. Because of these real-time requirements, there are limitations on the resolution of the observations, the size of the data sets, and the sophistication of retrieval algorithms and forward models. Historically, the calibration and stability of operational satellites has been of secondary importance; they are essentially calibrated by the conventional data as communicated through the assimilation system.

- Research Satellite Data: Research satellite data are those instruments that have been launched for scientific exploration or technology proof of concept. Research satellites measure parameters that have not been measured before or measure parts of the environment that have been adequately sampled. For example, measurements of carbon dioxide from space would be a new, important measurement; measurements of water and temperature in the planetary boundary layer would resolve these parameters in a region of especial importance. Research satellites might also seek to resolve traditional satellite measurements, temperature or ozone, with higher accuracy, higher spectral resolution, or higher horizontal and vertical resolution than operational satellites.

The entire suite of observations must be considered in a comprehensive data assimilation activity. Even if a research group is considering the impact of a particular instrument on a particular problem, usually a foundational data assimilation system is required that adequately considers the primary variables from the conventional and operational satellite data systems.

Though small in number, in meteorological data assimilation the conventional data are very important to the quality of the analysis and the forecast. This is perhaps due to the high quality of these observations as well as a good knowledge of the error characteristics. Alternatively, the assimilation systems might be implicitly tuned to these observations because of their long heritage and the legacy of adding new data types onto the pre-existing data system. Scientific investigation and re-investigation of the existing conventional and operational data systems would be an interesting and potentially productive research activity.

By shear quantity, the operational satellite data far outnumber the conventional data. The satellite data assures high quality global analyses, and satellite data have been essential to the continual improvement of weather forecasts. Research satellite data not only improve the quality of analyses relative to the pre-existing data system, but extend the analysis to new parameters and new domains; for instance, constituents and chemistry, land-surface, ocean prediction, *etc.* Because of the great cost of research instruments, there is increasing use of research observations at operational centers to assure that the research instruments benefit society.

The use of research satellites in operation applications greatly increases the complexity of the assimilation process. For many years conventional data and operational satellites worked under tight conventions that assured both common data formats and a small number of data centers that supplied all of the nations of the world. However, with the use of research observations, centers and scientists must interact with many specialized data processing organizations and use a multitude of data formats. In addition, both the operational and research satellites are contributing to a vast increase in the number of available observations. This increase in the number of observations, projected to be as much as six orders of magnitude between the years 2000 and 2010, will overwhelm data systems and computational capabilities unless new techniques are developed for the selection and use of data to be assimilated. Data usage provides, yet another, major challenge and is the focus of much of the research and development at operational centers.

An accurate operational data assimilation system provides the ideal interface between scientists and observing systems. There are many possibilities for developing an adaptive observing system – *i.e.* a tunable sensor web. In fact, such a notion provides one of the strategies for addressing the computational challenges of the projected enormous increase in data volume. The data assimilation system could target which of the observations from the satellites might be expected to have the greatest impact on a particular application, *e.g.* the forecast. Then only those data might be selected for evaluation or retrieval and possible inclusion in the data assimilation. In the future, it might be possible to direct the satellite to particular parts of the Earth and to target and take, only, those observations expected to have the greatest impact.

There are two natural places for the data assimilation system to provide interfaces to an adaptive observing system (see Figure 5). The first is the forecast, where particular features might be identified several hours or days in advance, then targeted. The second is with the observation minus forecast (O-F) increments. Large increments indicate places where the expected values from the forecast agree poorly with the observations. There are many possible reasons for disagreement, and one possibility is a region of high uncertainty, perhaps due to a poorly simulated developing system. Extra retrievals or targeted observations from any data platform could verify or refute the existence of such a developing system.

Finally, there are two basic strategies of observing. One is targeted observing of features or processes of special interest. The other is sampling or surveying of the entire domain. Both of these strategies are essential ingredients of scientific investigation. It is not a matter of one or the other. Robust assimilation depends on the existence of an observing system that adequately samples the domain. With this foundation, the idea of targeted observations to investigate those features that are not adequately represented by routine sampling makes sense.

High-performance Computing

Two aspects of computing will be discussed. First, the characteristics of the computational problem that distinguish Earth-science modeling and assimilation will be discussed. Second, the attributes that influence the increasing need for computational resources will be discussed.

The dictionary definition of a computer is:

- A device that determines a result by mathematical or logical operations.

High-performance computing describes a niche of computing that is associated with those platforms able to address the most demanding calculations. Since all aspects of computational technology (processor speed, memory, storage, network speed, *etc.*) show an exponential increase in capability as a function of time, the technical specifications of high-performance computing is also a function of time. High-performance computing is also called supercomputing and high-end computing [NAS, 2005].

There are a number of potential definitions or descriptions for high performance computing. These include:

- Computing that is 1 – 2 orders of magnitude beyond that available from the state-of-the-art desktop environment, or alternatively, beyond that which can be acquired by a well funded principal investigator.

- "The class of fastest and most powerful computers available," from Landau and Fink [1993]

There are two important attributes which are common to applications requiring high-performance computers. The first is that multiple computational processors must be gathered together and made to operate on a single image of the application software in order to achieve acceptable time to solution. Current practices in high-performance computing centers would suggest that applications that require approximately 64 processors on a single job would be termed high-performance. The second attribute is that special attention must be paid to the management of memory during the run time of the application.

These attributes highlight that high-performance computing is not simply an issue of hardware, but one also of software. In order to make effective use of a high-performance computer the scientist must have high-performance software. High-performance software must be able to scale to multiple processors; that is, the software must be able to utilize additional processors efficiently. As additional processors are added, the efficiency of each added processor is reduced because of communications overhead. There is a point at which adding more processors does little to increase the performance of the application. At the heart of efficient scaling is the management of memory. If the information needed by the processors can be kept at ready access to the processors, then efficient scaling can be maintained. This suggests that memory bandwidth; that is, how fast does information transfer from memory to the processor is an important aspect of a high-performance computer. In many applications that involve fluid flow, the physics of the problem require that information from one processor be communicated to other processors. This provides a formidable challenge to the scientific programmer, which is specifically related to the memory architecture of the hardware. This brings the need to add specialized computer programmers to the teams that aspire to comprehensive modeling and data assimilation activities. NAS [2001] provides an excellent examination of the problems writing scaleable software for climate models and the interaction of hardware and software.

There are two competing approaches by computer manufacturers to build high-performance computers. The first is to build specialized platforms that are anchored around custom processors, custom memory architecture, and custom communication interconnects amongst the processors. This adds significant cost to the computational platform. Since high-performance computing is a small part of the market, these tightly integrated platforms do not provide cost-performance numbers that appeal to the majority of the market. The second strategy, therefore, is to build high-performance computers out of components that are commercially available. This takes advantage of the exponential growth of increasing component capability. However, this requires that the computer companies build the environment that connects these components together – components that have not generally been developed to work together. This, again, adds significant cost to the computational platform. Further, this second approach pushes more of the work to the scientific programmer developing high-performance software. The issues of high-performance computing, its role in science, and their link to market factors are discussed in NAS [2005].

The need for high-performance computing is driven by both the requirement that scientific investigation requires a certain level of computational completeness to be productive and the requirement that the time to solution allows the products of the computation to be useful in their application. The workload suggested by these requirements falls into two natural categories – capability computing and capacity computing. These are described in Figure 8. Capability is defined by the maximum number of processors that can be focused, efficiently, on a single application. Capability is generally driven by a demand for increasing realism and comprehensiveness in a calculation, or a requirement that a product be produced in a given time segment (*i.e.,* real-time requirements). Capacity generally describes the execution of many applications that individually do not require highest capability. An example of capability computing would be a high-resolution, deterministic weather forecast; an example of capacity computing would be an ensemble of low-resolution forecasts to develop probabilistic information. Both capability and capacity computing are important to Earth-science modeling and assimilation. Sometimes, however, scientists are limited in capability experiments because of the expense and difficulty of writing high-performance software.

A heuristic example that demonstrates the communication issues of high-performance computing can be made from consideration of a group of people who need to make a series of transactions. If each person can make their transaction without negotiation with other people, for instance, buying their own lunch, then a group of people is well served by having a number of cashiers. However, if the group ordered their lunches together and need to negotiate with each other over the amount that each individual needs to pay, and further, requires the cashier to participate in the execution of their negotiation, then having more than one cashier is of little benefit. The computational problem is, therefore, defined not only by the number of transactions (calculations), but also by the amount of negotiation (communications) required. Figure 9 uses the format of Figures 2 and 3 to illustrate this point for the modeling of a hurricane. Assuming that the grid points are now associated with a certain subset of the processors (Figure 8), then information from one processor is needed from other processors to determine the internal dynamics of the hurricane. In addition, the path that the hurricane follows connects information from a series of grid points and processors. In is intuitive that the choice of grid, the specification of variables, and the selection of a discretization routine will impact the computational performance. While this example demonstrates the need for communications in a particular problem, other applications, for example land-surface assimilation, might only have weak requirements for communications. Therefore, loosely connected computation platforms might be adequate.

Finally, there are several aspects of the modeling and assimilation problem that stress computational systems and push capability requirements. The common ones in modeling are increased resolution, improved physics, inclusion of new processes, and integration and concurrent execution of Earth-system components that are normally run separately – that is, coupled models. Often, real-time needs define capability requirements. When considering data assimilation the computational requirements become much more challenging. Often the computational characteristics of the statistical analysis are defined as a function of the number and distribution of the observations; therefore, the increasing number of observations could be computationally crippling. Advanced assimilation techniques often involve iterative cycling between the model and the statistical analysis routine, increasing the computational burden. The increasing diversity of data sources and the use of research observations in assimilation place tremendous demands on networks and data systems. The computational details of modeling, statistical analysis, and quality control are quite different. As with the construction of a state-of-the art model or data assimilation system, balanced cost considerations need to be made in the computational aspects of the problem – both software and hardware. This requires the scientist and the science manager to constantly consider the tension between the reduction of the problem to its component parts and the unification of those parts into a system.

Summary

This lecture introduced the fundamental ideas that a scientist needs to understand when building or using models in Earth-science research. Rather than focusing on technical aspects of modeling and data assimilation, the lecture focused on a number of underlying principles. These principles, if adhered to, will allow the models and model products to be used in quantitative, data-driven research. This goes beyond simple comparison of models and observations and using their similarity as a measure of worth.

With regards to stand-alone models in the absence of data assimilation, it was emphasized that the underlying physics should be well represented. This requires special attention to the physics as using accurate numerical techniques does not guarantee physical consistency. Data assimilation was introduced as adding a forcing term to the model that is a correction based on observations. This additional forcing term changes the balance of forces. Therefore, budgets calculated from assimilated data are not expected to be robust for geophysical applications. For both modeling and data assimilation, it is much more difficult to quantitatively analyze and interpret the results than it is to develop new modules and components. Few scientists do this analysis well, and students are challenged to learn existing techniques and to develop new techniques.

With regard to data assimilation, the importance of the observing system was emphasized. This requires monitoring of the observing system and vigorous attention to quality control. It also requires attention to the details of the instrumentation, for example, the observational technique. Scientific investigation of the observing system was encouraged. The importance of bias in data assimilation was also discussed. The presence of bias lies at the foundation of the physical consistency of assimilated data sets. While data assimilation has had a number of outstanding successes, these issues of bias and physical consistency require scientists to consider the appropriateness of data assimilation to their particular problem.

The attention to the observing system brought out the changing nature of the observing system. Specifically, the observing system is becoming more diverse and data volumes are increasing rapidly. This requires the efforts of many scientists and new computational techniques to utilize these new observations effectively. Data assimilation systems provide a natural link between scientists and the observing system, including the possibility of adaptive observing systems.

Finally, the computational aspects of modeling and assimilation were discussed. Comprehensive activities that address the entire Earth system remain beyond the most capable computers. Special challenges come from the fact that many computations are required (transaction) and communication is required between the computations (negotiation). In addition the computational issues of faced when embracing the data systems are often different than those usually considered in stand-alone modeling activities. Computational considerations must be incorporated in the development of data assimilation systems, and again, they need to also address issues of physical consistency.

The end-to-end data assimilation system must have balance. There is little benefit developing components to a high state of accuracy or performance if there are other weaknesses in the system. Experience suggests that future progress will be most effectively realized through the use of new data types and improving the representation of parameterized physics in models. With these efforts, bias might be addressed in a fundamental way. If the bias problem is not addressed, then there are intrinsic limitations to the problems appropriately addressed by data assimilation.

References

Brasseur, G., and S. Solomon, *Aeronomy of the Middle Atmosphere,* D. Reidel Publishing Company, 452 pp, 1986.

Cohn, S. E., An introduction to estimation theory, *J. Met. Soc. Japan,* **75 (1B)**, 257-288, 1997.

Daley, R., *Atmospheric Data Analysis,* Cambridge University Press, 457 pp, 1991.

Dee, D. P., and A. da Silva, Data assimilation in the presence of forecast bias, *Q. J. Roy. Meteor. Soc.,* **124** (545), 269-295 Part A, 1998.

Dee, D. P., L. Rukhovets, R. Todling, A. M. da Silva, and J. W. Larson, An adaptive buddy check for observational quality control, *Q. J. Roy. Meteor. Soc.,* **127** (577), 2451-2471 Part A, 2001.

Dessler, A. E., *The Chemistry and Physics of Stratospheric Ozone*, Academic Press, 214 pp, 2000.

Douglass, A. R., M. R. Schoeberl, R. B. Rood, and S. Pawson, Evaluation of Transport in the Lower Tropical Stratosphere in a Global Chemistry and Transport Model, *J. Geophys. Res.*, **108**, Art. No. 4259, 2003.

Holton, J. R., *An Introduction to Dynamic Meteorology,* Elsevier Academic Press, 535 pp, 2004.

Jacobson, M. Z., *Fundamentals of Atmospheric Modeling,* Cambridge University Press, 672 pp, 1998 - (2nd Ed, to appear, 2005).

Jackman C. H., P. D. Guthrie, and J. A. Kaye, An intercomparison of nitrogen-containing species in Nimbus 7 LIMS and SAMS data, *J. Geophys. Res.,* **92**, 995-1008, 1987.

Johnson, S. D., D. S. Battisti, and E. S. Sarachik, Empirically derived Markov models and prediction of tropical Pacific sea surface temperature anomalies, *J. Climate,* **13**, 3-17, 2000.

Kistler R., E. Kalnay, W. Collins, *et al.,* The NCEP-NCAR 50-year Reanalysis: Monthly means CD-ROM and documentation, *Bull. Amer. Meteor. Soc.* **82**, 247-267, 2001.

Lait, L. R., Systematic differences between radiosonde measurements, *Geophys. Res. Lett.,* **29**, Art. No. 1382, 2002.

Landau, R. H., and P. J. Fink, *A Scientist's and Engineer's Guide to Workstations and Supercomputers,* John Wiley and Sons, 416 pp, 1993.

Lin, S. J., A “vertically Lagrangian” finite-volume dynamical core for global models, *Mon. Wea. Rev.,* **132**, 2293-2307, 2004.

Long, C. S., A. J. Miller, H. T. Lee, J. D. Wild, R. C. Przywarty, and D. Hufford, Ultraviolet index forecasts issued by the National Weather Service, *Bull. Amer. Meteorol. Soc.*, **77**, 729-748, 1996.

Molod, A., H. M. Helfand, and L. L. Takacs, The climatology of parameterized physical processes in the GEOS-1 GCM and their impact on the GEOS-1 data assimilation system, *J. Climate,* **9**, 764-785, 1996.

NAS, *Improving the Effectiveness of U.S. Climate Modeling*, National Academy Press, Washington, DC, 142 pp, 2001 (http://books.nap.edu/catalog/10087.html).

NAS, *Getting Up to Speed: The Future of Supercomputing,* National Academy Press, S. L. Graham, M. Snir, and C. A. Paterson (Eds.), Washington, DC, 308 pp, 2005 (http://books.nap.edu/catalog/11148.html).

Plumb, R. A., and M. K. W. Ko, Interrelationships between mixing ratios of long lived stratospheric constituents, *J. Geophys. Res.,* **97**, 10145-10156, 1992.

Randall, D. A. (Ed.), *General Circulation Model Development: Past, Present, and Future,* Academic Press, 807 pp, 2000.

Rood, R. B., Numerical advection algorithms and their role in atmospheric transport and chemistry models, *Rev. Geophys.,* **25**, 71-100, 1987.

Schoeberl, M. R., A. R. Douglass, Z. Zhu, and S. Pawson, A comparison of the lower stratospheric age-spectra derived from a general circulation model and two data assimilation systems, *J. Geophys. Res.,* **108**, Art. No. 4113, 2003

Stajner, I., L. P. Riishojgaard, and R. B. Rood, The GEOS ozone data assimilation system: Specification of error statistics, *Q. J. R. Meteorol. Soc.,* **127**, 1069-1094, 2001.

Stajner, I., N. Winslow, R. B. Rood, and S. Pawson, Monitoring of Observation Errors in the Assimilation of Satellite Ozone Data, *J. Geophys. Res.,* **109**, Art. No. D06309, 2004.

Swinbank, R., R., Shutyaev, and W. A. Lahoz (Eds.), *Data Assimilation for the Earth System,* NATO Science Series: IV: Earth and Environmental Sciences, **26,** Kluwer, 388 p, 2003.

Trenberth, K. E. (Ed.), *Climate System Modeling,* Cambridge University Press, 788 pp, 1992.

Wargan, K., I. Stajner, S. Pawson, R. B. Rood, and W. –W. Tan, Assimilation of Ozone Data from the Michelson Interferometer for Passive Atmospheric Sounding, *Quart. J. Royal. Met. Soc.*, to appear, 2005.

FIGURES: Fundamentals of Modeling, Data Assimilation, and (High Performance) Computing

Richard B. Rood
Chief, Computational and Information Science and Technology Office

NASA/Goddard Space Flight System

Figure 1: Simulation Framework

(General Circulation Model, “Forecast”)

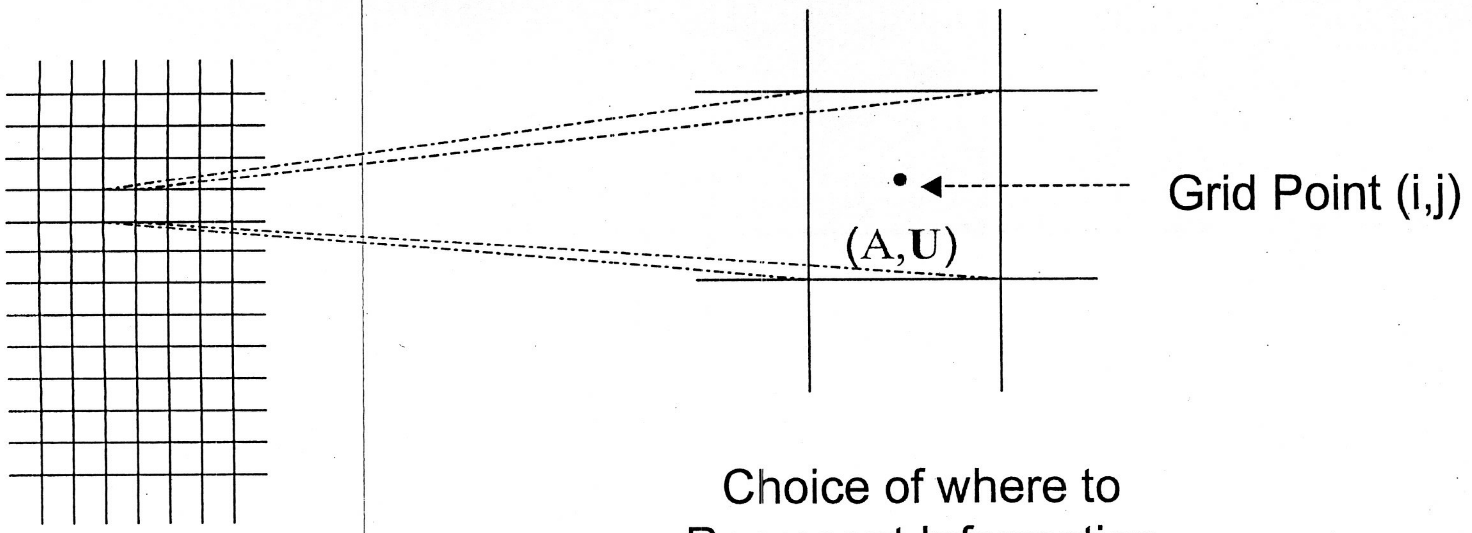
Boundary Conditions	Emissions, SST, …	ε
Discrete/Parameterize	$(A_{n+\Delta t} - A_n)/\Delta t = \ldots$	$(\varepsilon_d, \varepsilon_p)$
Theory/Constraints	$\partial u_g/\partial z = -(\partial T/\partial y)R/(Hf_0)$	Scale Analysis
Primary Products (*i.e.* A)	$T, u, v, \Phi, H_2O, O_3 \ldots$	$(\varepsilon_b, \varepsilon_v)$
Derived Products (F(A))	Pot. Vorticity, v*, w*, …	Consistent

$(\varepsilon_b, \varepsilon_v)$ = (bias error, variability error)

Derived Products likely to be physically consistent, but to have significant errors. *i.e.* The theory-based constraints are met.

Figure 2: Discretization of Resolved Transport

- $\partial A/\partial t = - \nabla \bullet \mathbf{U}A$

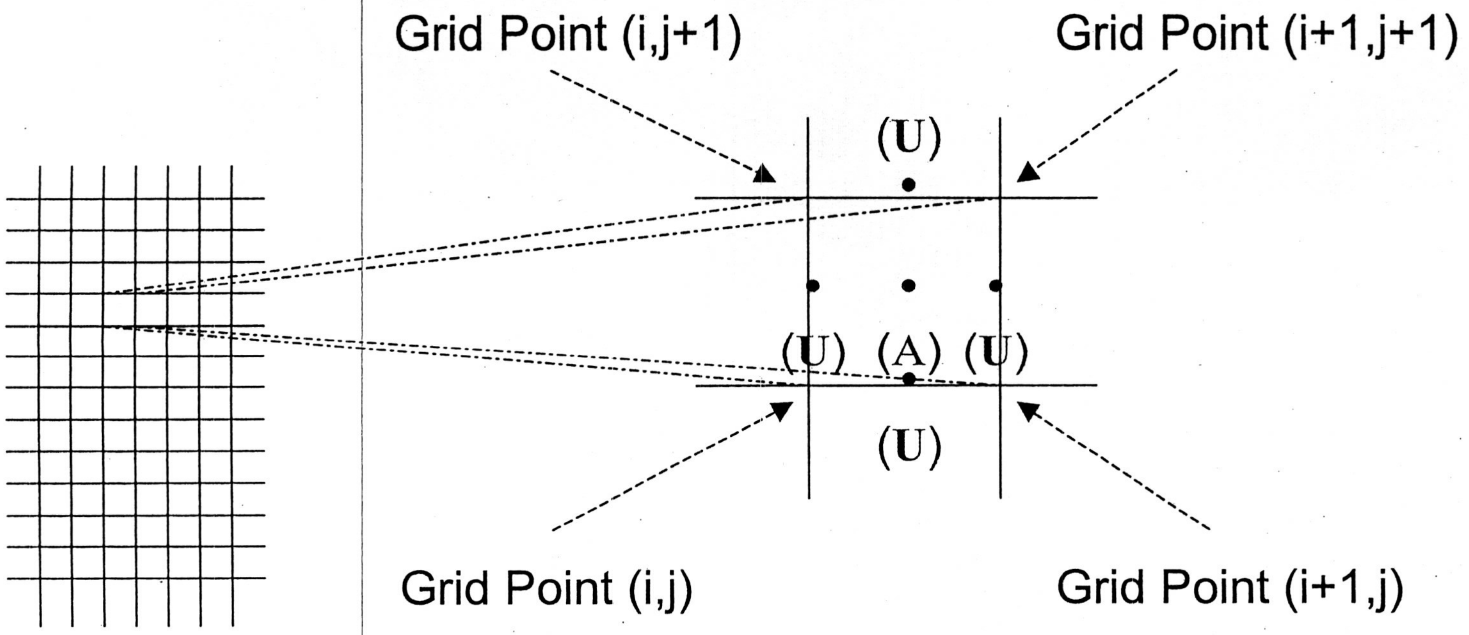


Gridded Approach
Orthogonal?
Uniform area?
Adaptive?
Unstructured?

Choice of where to
Represent Information

Choice of technique to
approximate operations in
representative equations
Rood (1987, Rev. Geophys.)

Figure 3: Discretization of Resolved Transport



<u>Choice of where to Represent Information Impacts Physics</u>

- Conservation
- Scale Analysis Limits
- Stability

Figure 4: Assimilation Framework

Model Data

Emissions, SST, …	ε	Boundary Conditions
$(A_{n+\Delta t} - A_n)/\Delta t = \ldots$	ε	Discrete/Error Modeling
$\partial u_g/\partial z = -(\partial T/\partial y)R/(Hf_0)$	Scale Analysis	Constraints on Increments
$A_i \equiv T, u, v, \Phi, H_2O, O_3 \ldots$	$(\varepsilon_b, \varepsilon_v)$	$(\varepsilon_b, \varepsilon_v)$ reduced
Pot. Vorticity, v*, w*, …	Consistent	Inconsistent

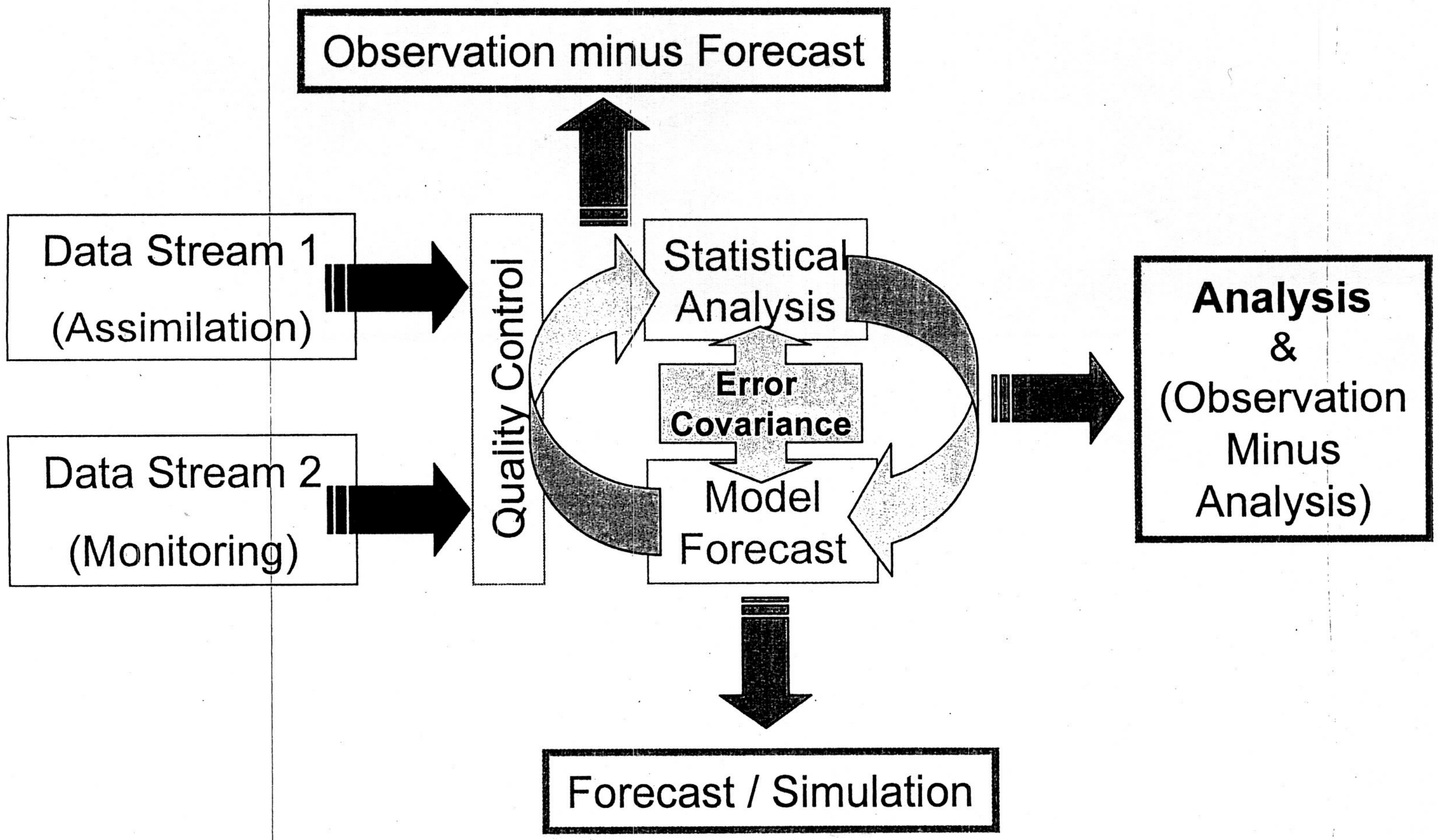
O is the "observation" operator; P^f is forecast model error covariance R is the observation error covariance; x is the innovation

Generally assimilate resolved, predicted variables. Future, assimilate or constrain parameterizations. (T, u, v, H2O, O3)

Data appear as a forcing to the equation

Does the average of this added forcing equal zero?

Figure 5: Schematic of Data Assimilation System



Observation minus Forecast

Data Stream 1
(Assimilation)

Data Stream 2
(Monitoring)

Quality Control

Statistical Analysis

Error Covariance

Model Forecast

Analysis
&
(Observation
Minus
Analysis)

Forecast / Simulation

Figure 6: Quality Control: Interface to the Observations

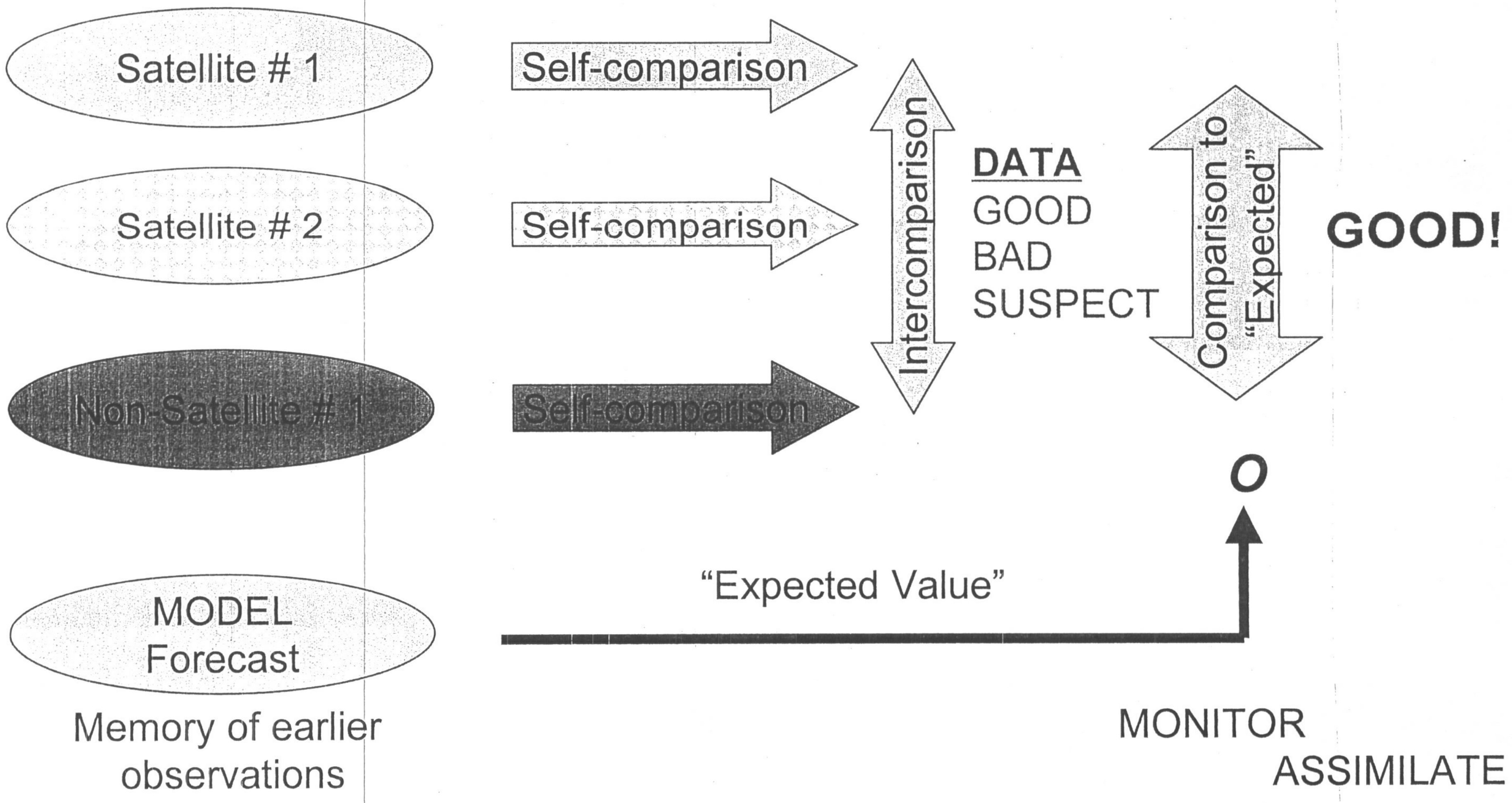
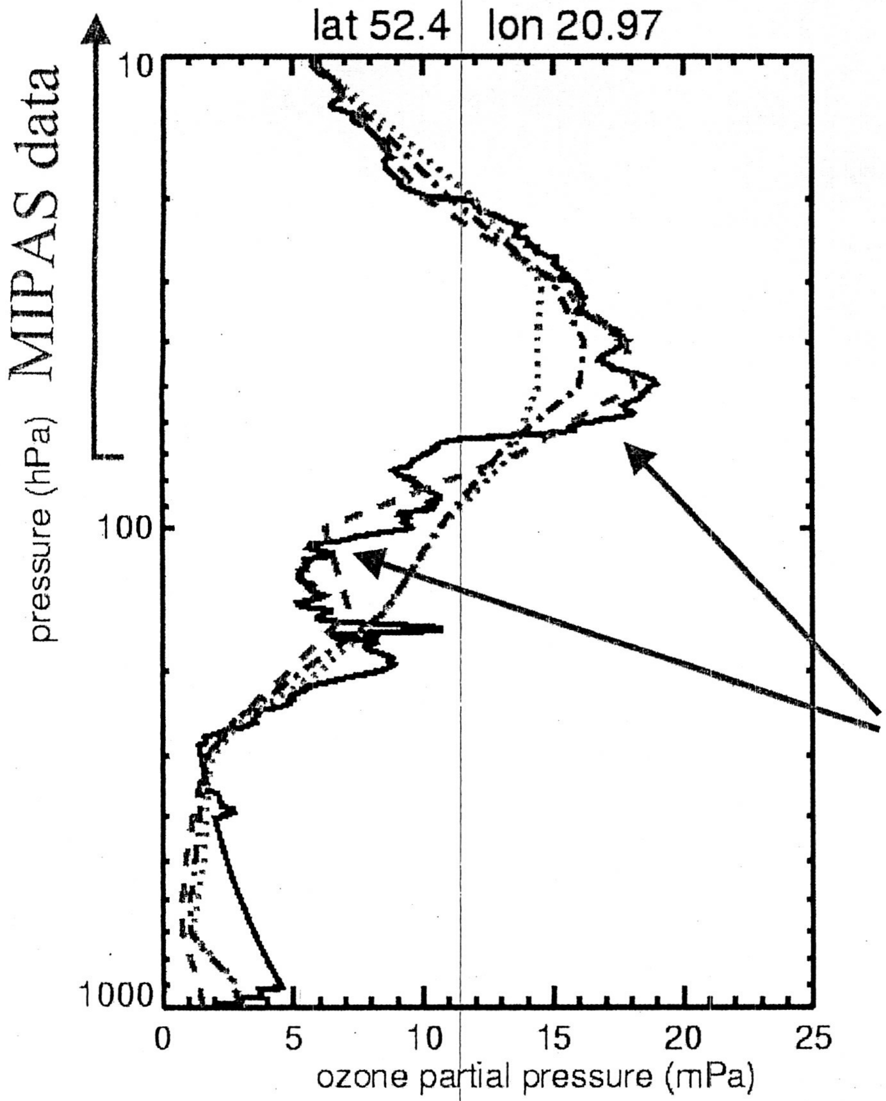


Figure 7: MIPAS ozone assimilation



- Comparison of an individual ozonesonde profile with three assimilations that use <u>SBUV total column</u> and stratospheric profiles from:
  - **SBUV**
  - **SBUV and MIPAS**
  - **MIPAS**
- MIPAS assimilation captures vertical gradients in the lower stratosphere
- Model + Data capture synoptic variability and spreads MIPAS information

Figure 8: Computational Capacity & Capability

- Capability: Execution in a given amount of time of a job requiring the entire computer.
  - Capability limit is the most challenging
- Capacity: Aggregate of all jobs running simultaneously on the computer.

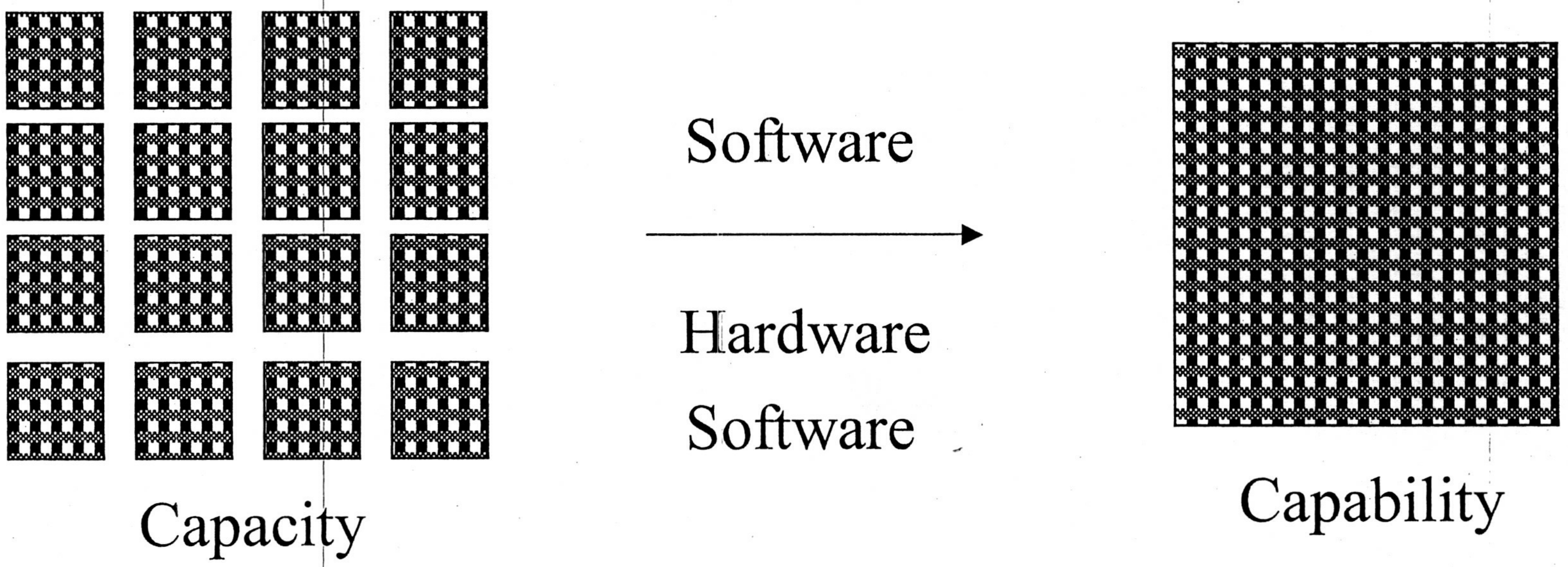
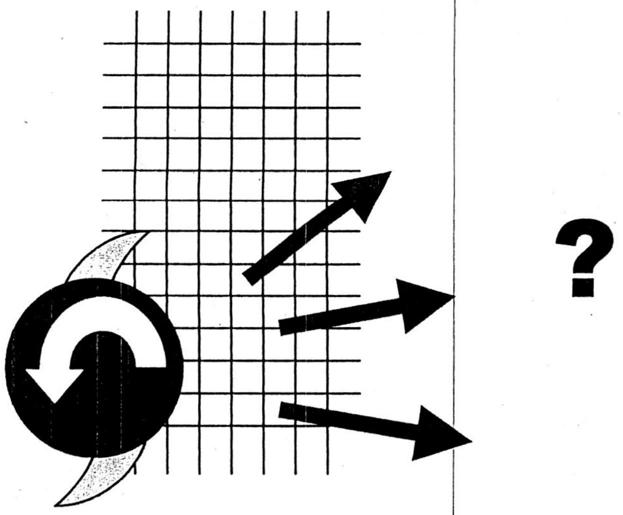
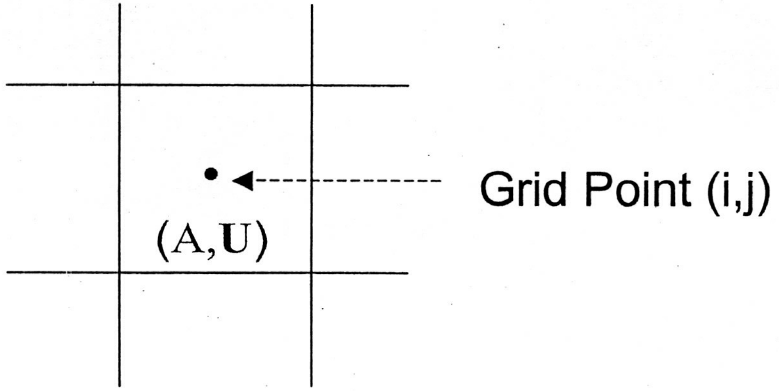


Figure 9: Requirement for parallel communication



Gridded Approach

Choice of grid impacts computational efficiency



Choice of where to represent information impacts computational efficiency